CUNY SPS DATA 621 - CTG5 - HW5

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

May 15th, 2019

Contents

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET	Number of Cases Purchased	count response
AcidIndex	Method of testing total acidity by using a weighted avg	continuous numerical predictor
Alcohol	Alcohol Content	continuous numerical predictor
Chlorides	Chloride content of wine	continuous numerical predictor
CitricAcid	Citric Acid Content	continuous numerical predictor
Density	Density of Wine	continuous numerical predictor
FixedAcidity	Fixed Acidity of Wine	continuous numerical predictor
FreeSulfurDioxide	Sulfur Dioxide content of wine	continuous numerical predictor
LabelAppeal	Marketing Score indicating the appeal of label design	continuous numerical predictor
ResidualSugar	Residual Sugar of wine	continuous numerical predictor
STARS	Wine rating by a team of experts. 4 = Excellent, 1 = Poor	continuous numerical predictor
Sulphates	Sulfate conten of wine	continuous numerical predictor
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	continuous numerical predictor
VolatileAcidity	Volatile Acid content of wine	continuous numerical predictor
pH	pH of wine	continuous numerical predictor

Table 2: Summary statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
TARGET	1	12795	3.03	1.93	3.00	3.05	1.48	0.00	8.0	8.00	-0.33
FixedAcidity	2	12795	7.08	6.32	6.90	7.07	3.26	-18.10	34.4	52.50	-0.02
VolatileAcidity	3	12795	0.32	0.78	0.28	0.32	0.43	-2.79	3.7	6.47	0.02
CitricAcid	4	12795	0.31	0.86	0.31	0.31	0.42	-3.24	3.9	7.10	-0.05
ResidualSugar	5	12179	5.42	33.75	3.90	5.58	15.72	-127.80	141.2	268.95	-0.05
Chlorides	6	12157	0.05	0.32	0.05	0.05	0.13	-1.17	1.4	2.52	0.03
FreeSulfurDioxide	7	12148	30.85	148.71	30.00	30.93	56.34	-555.00	623.0	1178.00	0.01
TotalSulfurDioxide	8	12113	120.71	231.91	123.00	120.89	134.92	-823.00	1057.0	1880.00	-0.01
Density	9	12795	0.99	0.03	0.99	0.99	0.01	0.89	1.1	0.21	-0.02
pH	10	12400	3.21	0.68	3.20	3.21	0.39	0.48	6.1	5.65	0.04
Sulphates	11	11585	0.53	0.93	0.50	0.53	0.44	-3.13	4.2	7.37	0.01
Alcohol	12	12142	10.49	3.73	10.40	10.50	2.37	-4.70	26.5	31.20	-0.03
LabelAppeal	13	12795	-0.01	0.89	0.00	-0.01	1.48	-2.00	2.0	4.00	0.01
AcidIndex	14	12795	7.77	1.32	8.00	7.64	1.48	4.00	17.0	13.00	1.65
STARS	15	9436	2.04	0.90	2.00	1.97	1.48	1.00	4.0	3.00	0.45

1 DATA EXPLORATION

1.1 Summary Statistics

1.2 Variable Descriptions

1.2.1 Summary Statistics Graphs

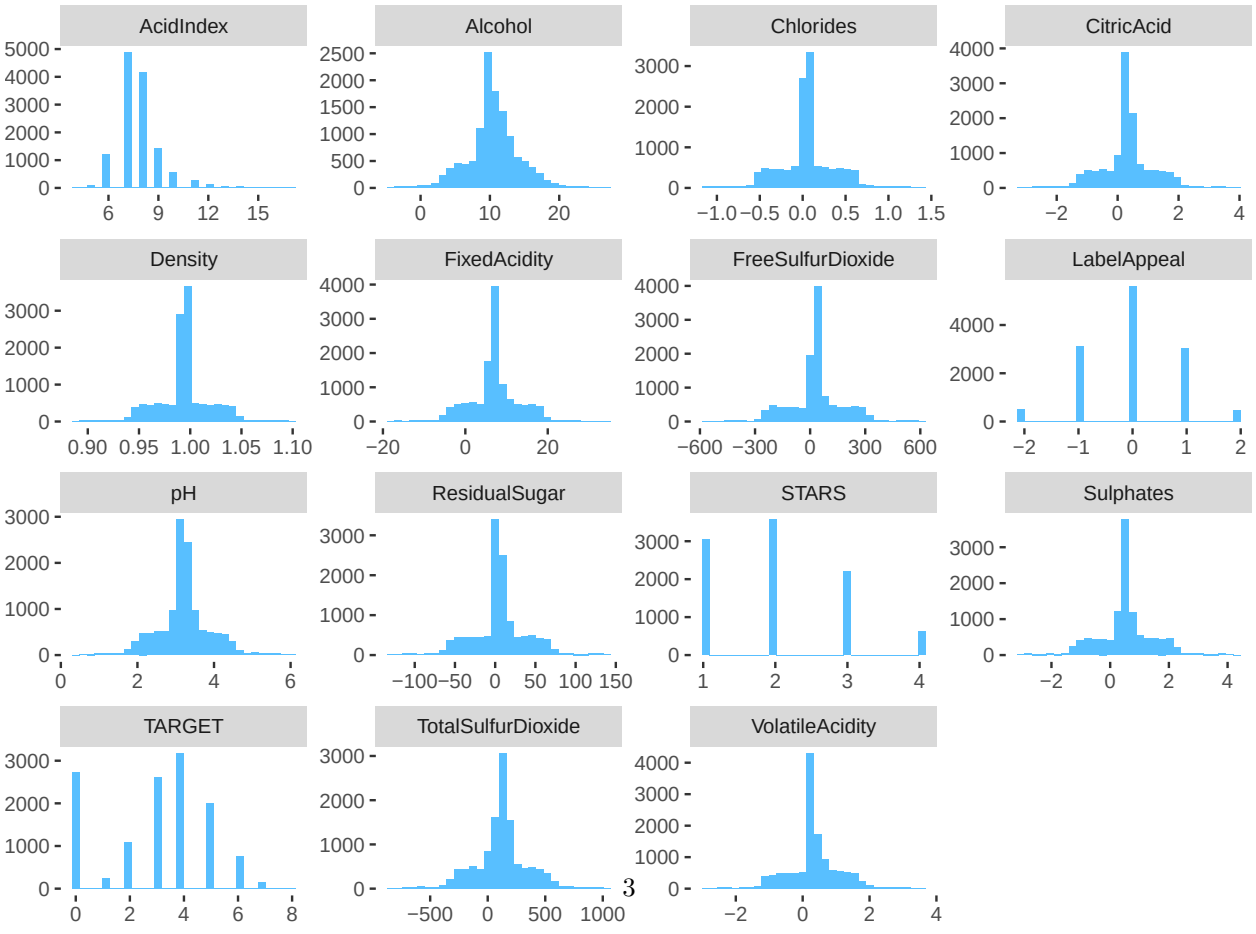

Figure 1: Numeric Data Distributions as a Function of TARGET_FLAG

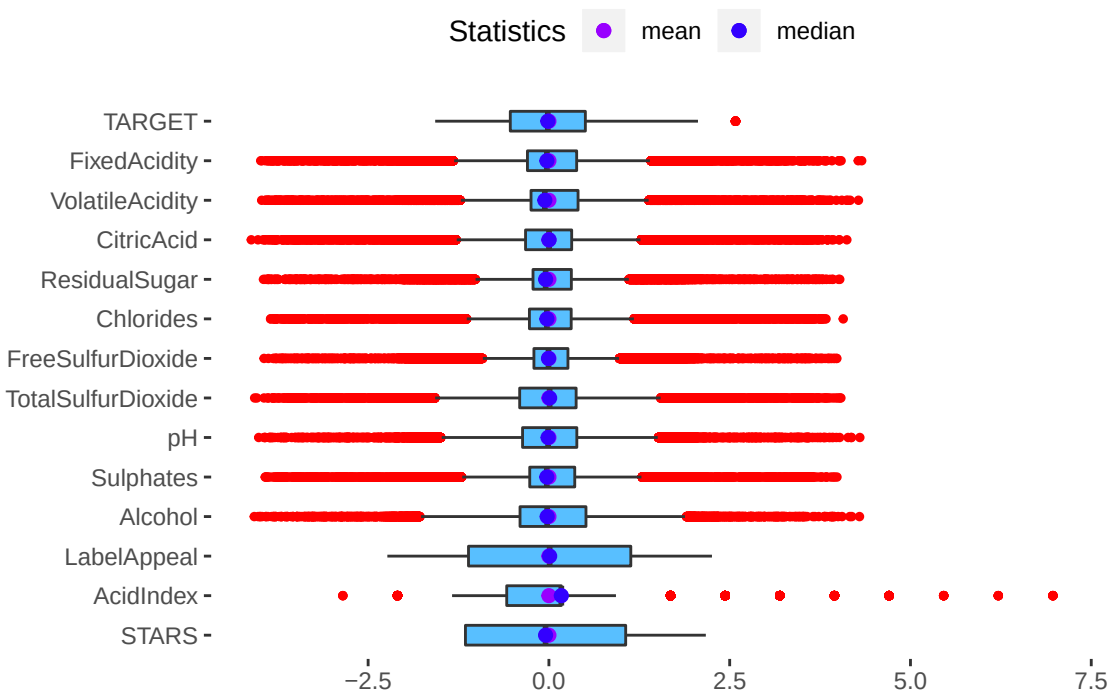

Figure 2: Scaled Boxplots

1.3 Linearity

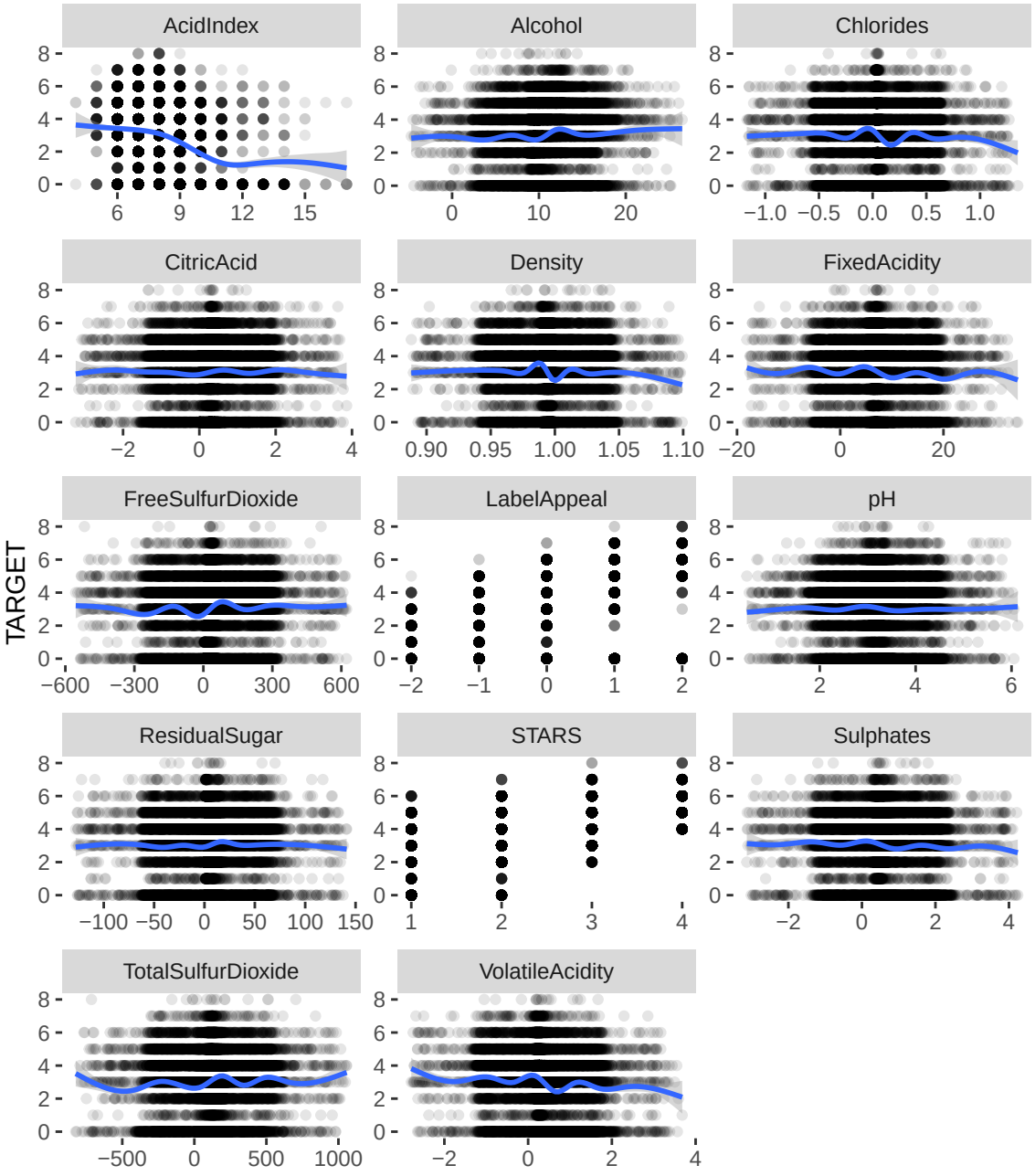

Figure 3: Scatter plot between numeric predictors and the TARGET

1.3.1 Log Transformed Data

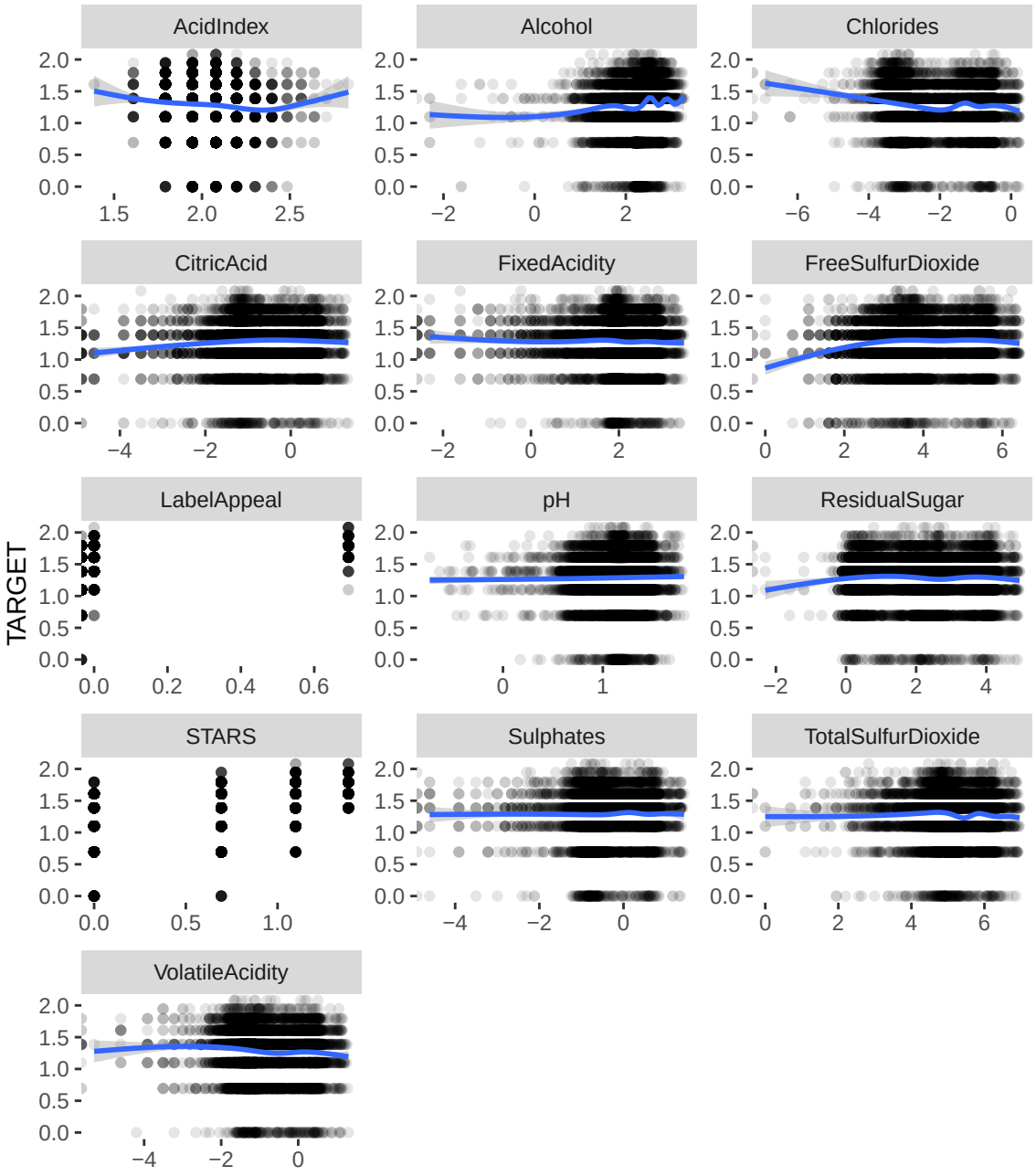

Figure 4: Scatter plot between log transformed predictors and the log transformed TARGET filtered for rows where TARGET is greater than 0

1.3.2 Box-Cox

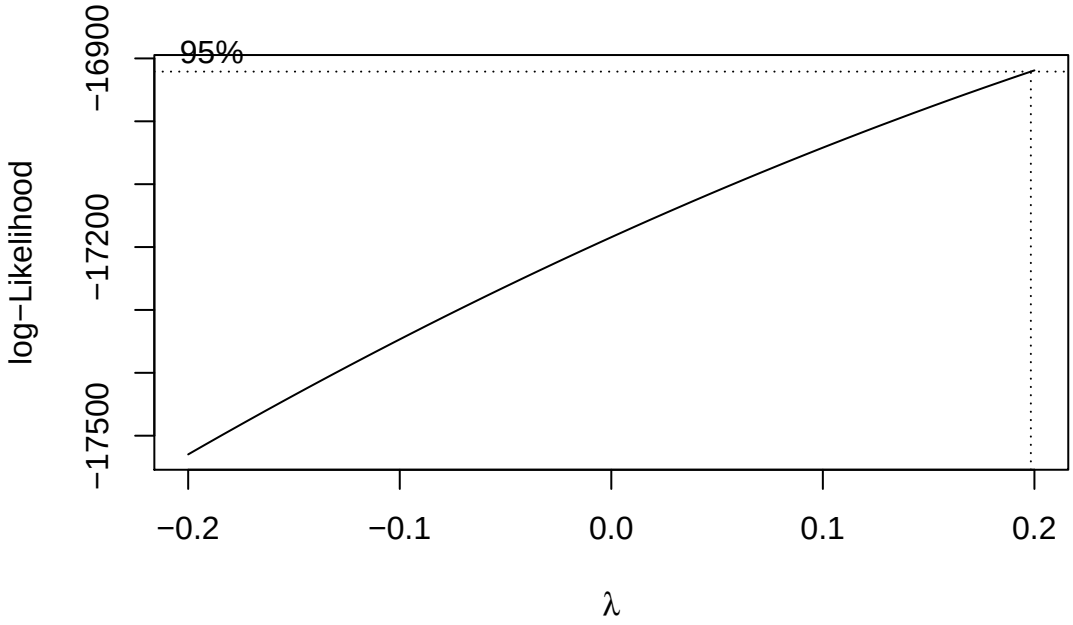

Figure 5: Box-Cox Plot

1.3.3 Square Root Transformed Predictors and Log Transformed Target

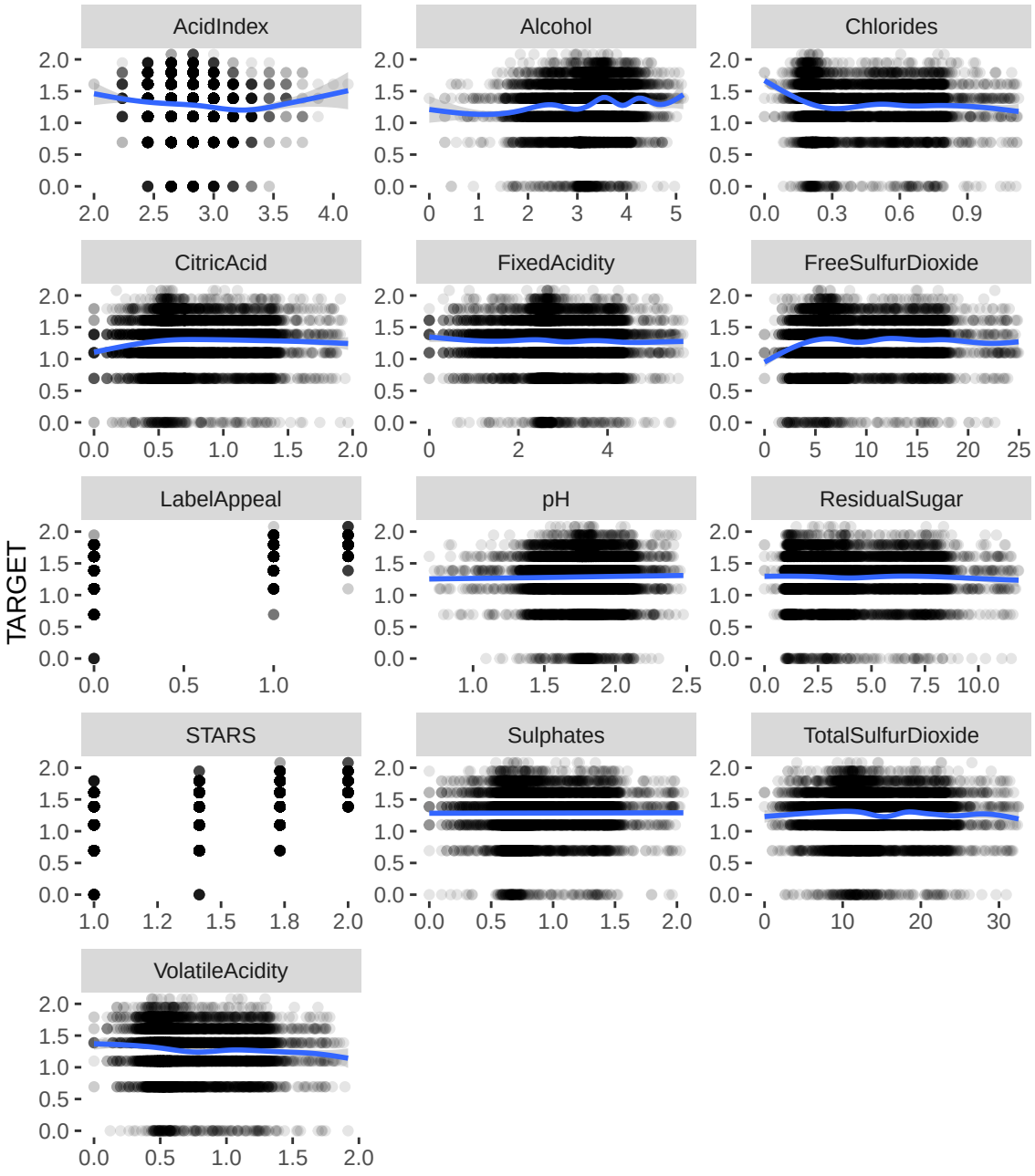

Figure 6: Scatter plot between square root transformed predictors and the square root transformed TARGET filtered for rows where TARGET is greater than 0

1.4 Missing Data

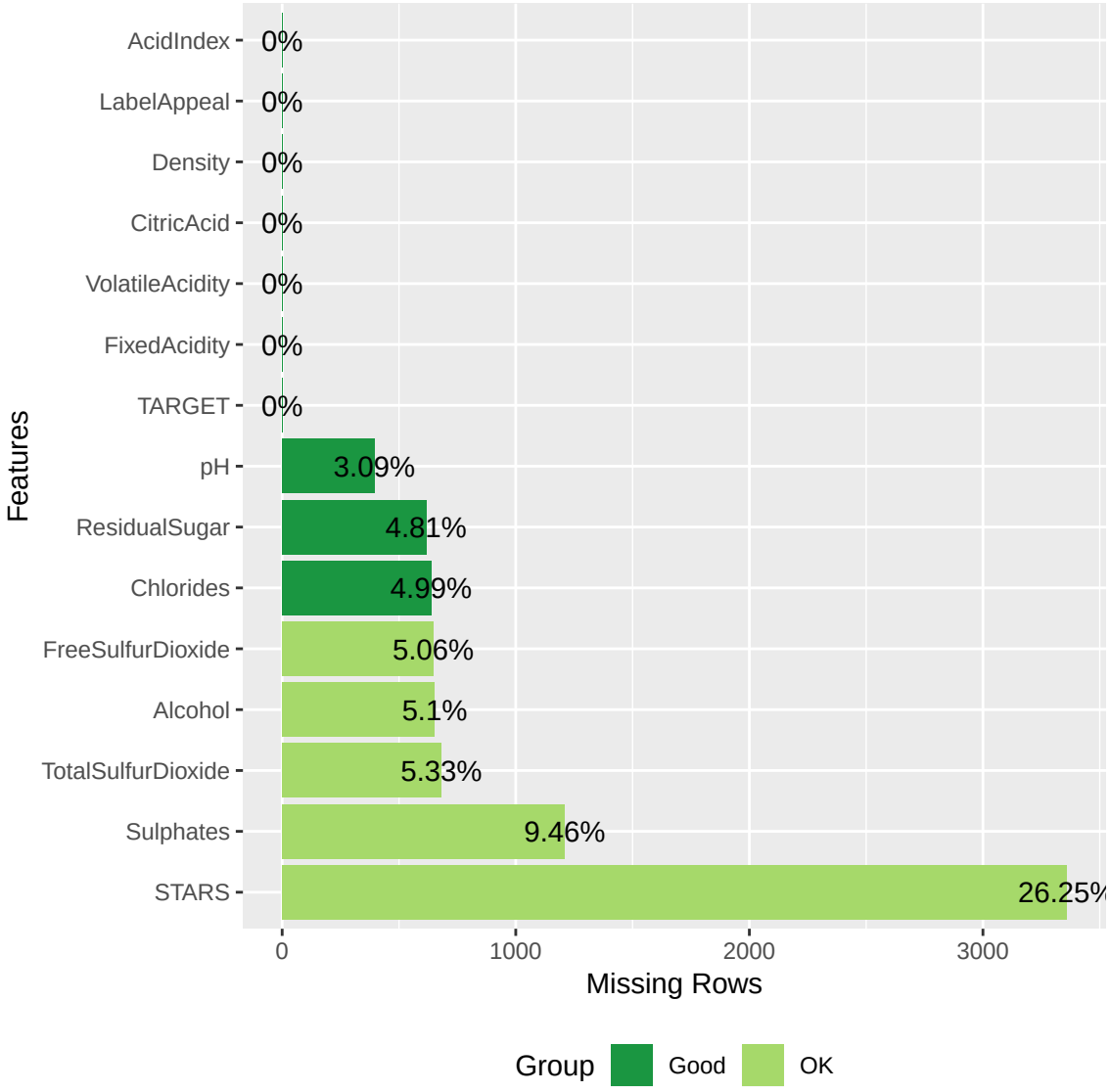

Figure 7: Missing data

2 DATA PREPARATION

Describe how you have transformed the data by changing the original variables or creating new variables

a. Fix missing values (maybe with a Mean or Median value)

b. Create flags to suggest if a variable was missing

c. Transform data by putting it into buckets

d. Mathematical transforms such as log or square root (or use Box-Cox)

e. Combine variables (such as ratios or adding or multiplying) to create new variables

2.1 Missing Values

3 BUILD MODELS

Using the training data set, build at least two different poisson regression models, at least two diffe

Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can

4 SELECT MODELS

Decide on the criteria for selecting the best count regression model. Will you select models with sligh

For the count regression model, will you use a metric such as AIC, average squared error, etc.? Be sure

4.1 Prediction

5 Appendix

The appendix is available as script.R file in `project5_wine` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining